Jurnal TEKNO KOMPAK, Vol. 14, No. 1, 2020, 5-10. P - ISSN 1412-9663, E-ISSN 2656-3525

PENENTUAN REKOMENDASI PELATIHAN PENGEMBANGAN DIRI BAGI PEGAWAI NEGERI SIPIL MENGGUNAKAN ALGORITMA C4.5 DENGAN PRINCIPAL COMPONENT ANALYSIS DAN DISKRITISASI

Hanif Rahmawan[1)], Azhari SN[2)]

[1)] Prodi S2 Ilmu Komputer, FMIPA UGM, Yogyakarta
[2)] Departemen Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
Bulaksumur, Caturtunggal, Depok, Sleman, D.I. Yogyakarta
Email : alyph3003@yahoo.com[1)], arisn@ugm.ac.id[2)]

Abstrak

Setiap institusi memiliki kebutuhan untuk terus meningkatkan pelayanan dan melakukan inovasi yang perlu mendapatkan dukungan dari SDM yang berkualitas. Pelatihan menjadi salah satu cara untuk mewujudkan SDM yang berkualitas. Namun terkadang penentuan pelatihan yang sesuai untuk seorang pegawai tidak mudah dan berpeluang menimbulkan ketidakkonsistenan. Masalah tersebut dapat diatasi dengan melakukan data mining terhadap data pemetaan pegawai sehingga didapatkan aturan untuk penentuan rekomendasi pelatihan pengembangan diri. Data pemetaan terdiri dari nilai aspek psikologis pegawai dan rekomendasi pelatihan yang diberikan oleh assessor.

Pada penelitian ini digunakan tiga metode, yaitu algoritma C4.5, kombinasi PCA, dan C4.5, serta kombinasi PCA, diskritisasi, dan C4.5 untuk melakukan penambangan pada data. Diskritisasi yang digunakan adalah diskritisasi berbasis entropi. Pada tahap pra-pemrosesan digunakan teknik over-sampling SMOTE untuk menangani 4 data pelatihan yang mengalami ketidakseimbangan kelas. Pada penerapan kombinasi algoritma PCA, diskritisasi, dan C4.5 dilakukan reduksi dimensi dengan menggunakan algoritma PCA. Data hasil reduksi didiskritisasi kemudian diklasifikasi dengan algoritma C4.5.

Hasil pengujian menunjukkan bahwa kombinasi PCA, diskritisasi, dan C4.5 memberikan performa yang lebih baik daripada kedua metode yang lain. Seluruh pelatihan menunjukan performa terbaik ketika diproses dengan metode ini. Penentuan rekomendasi pelatihan pengembangan diri bagi pegawai dapat dilakukan dengan metode ini dengan rerata akurasi 86,6%.

Kata kunci: *Pohon keputusan C4.5, pelatihan pengembangan diri, principal component analysis, diskritisasi berbasis entropi.*

1. Pendahuluan

Institusi baik negeri maupun swasta dituntut untuk lebih meningkatkan pelayanan dan senantiasa mengembangkan inovasi-inovasi baru. Dukungan dari sumber daya manusia (SDM) yang berkualitas mutlak diperlukan untuk dapat melakukan hal tersebut (Jantan et al. 2011). Salah satu cara untuk mendapatkan SDM yang berkualitas adalah dengan memberikan pelatihan terhadap SDM tersebut. Penentuan pelatihan yang sesuai untuk pegawai merupakan salah satu bentuk aktivitas manajemen talenta. Penentuan keputusan dalam manajemen talenta terkadang tidaklah mudah. Selain itu, penentuan keputusan juga bergantung pada berbagai macam faktor di antaranya faktor pengalaman, pengetahuan, preferensi, dan pertimbangan. Faktor-faktor tersebut dapat menyebabkan ketidakkonsistenan, ketidakakuratan, dan ketidaksamaan keputusan (Jantan et al. 2011).

Data mining merupakan sebuah metode untuk melakukan akuisisi pengetahuan. Dengan data mining, informasi-informasi implisit dan berharga dari sebuah data dapat diekstrak. Metode ini sudah digunakan di berbagai bidang, misal pada bidang keuangan untuk menentkan kelayakan calon debitur baru (Hermanto dan SN, 2017). Metode ini juga digunakan dalam bidang manajemen sumber daya manusia bahkan penggunaan dalam bidang tersebut meningkat cukup signifikan (Strohmeier & Piazza 2013).

Salah satu teknik data mining yang cukup populer adalah C4.5 (Hussain et al. 2013). Jantan et al. (2011) membandingkan algoritma C4.5 dengan beberapa algoritma lainnya dan hasilnya menunjukkan bahwa algoritma C4.5 cukup potensial digunakan di dalam bidang manajemen SDM.

Pengembangan algoritma C4.5 pun telah banyak dilakukan. Beberapa pengembangan berfokus pada pra pemrosesan yang dilakukan terhadap data sebelum diklasifikasi dengan algoritma C4.5. Hussain et al. (2013) mengembangkan algoritma C4.5 dengan cara melakukan pengurangan dimensi pada data yang akan diklasifikasi. Pengurangan data dilakukan dengan algoritma PCA. Hasil penelitiannya menunjukkan bahwa cara tersebut dapat meningkatkan efisiensi proses pembentukan pohon keputusan dan dapat meningkatkan akurasi dari proses klasifikasi.

Penggunaan algoritma PCA akan membuat data yang dihasilkan menjadi bersifat kontinu. Algoritma C4.5 dapat menangani atribut kontinu dengan melakukan diskritisasi lokal. Diskritisasi pada atribut kontinu sebelum dilakukan klasifikasi (diskritisasi global) akan dapat meningkatkan efisiensi proses klasifikasi bahkan dapat meningkatkan akurasi (Kareem dan Duaimi 2014). Hussain et al. (2013) dalam penelitiannya menggunakan diskritisasi lokal C4.5 sehingga pendekatan yang dilakukannya masih berpotensi untuk ditingkatkan lagi.

2. Pembahasan

2.1 Analisis Sistem

Data yang digunakan adalah data hasil pemetaan pegawai negeri sipil dari Badan Kepegawaian Negara. Data ini berisi hasil pemetaan pegawai yang dituangkan ke dalam 14 nilai aspek psikologis berikut ini:

1. Potensi kecerdasan
2. Daya konseptual
3. Fleksibilitas berpikir
4. Kemampuan numerik
5. Stres tolerance
6. Hasrat berprestasi
7. Sistematika kerja
8. Daya analisis
9. Stabilitas emosi
10. Kepercayaan diri
11. Penyesuaian diri
12. Inisiatif
13. Kerjasama
14. Kepemimpinan

Range dari nilai aspek psikologis tersebut adalah dari 0 sampai dengan 5 dengan masing-masing kemungkinan penambahan nilai + dan – untuk masing-masing level nilai, contoh 1-,1, dan 1+. Data tersebut juga disertai dengan rekomendasi pelatihan pengembangan diri bagi pegawai terkait. Penentuan rekomendasi pelatihan pengembangan diri dilakukan oleh assessor berdasar hasil tes pemetaan pegawai. Satu pegawai dapat diberikan lebih dari 1 rekomendasi pelatihan yang jenis pelatihannya sebagai berikut:

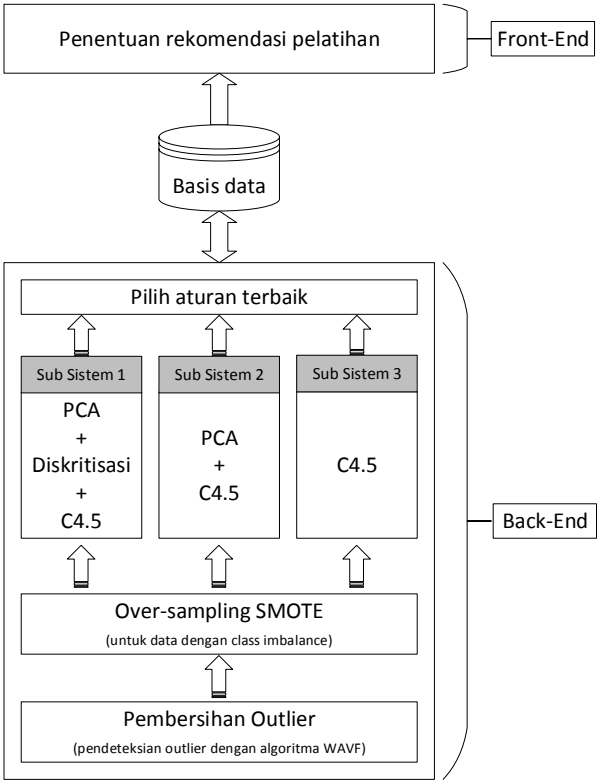
1. Achievement Motivation Training (AMT)
2. Effective Communication Skill (ECS)
3. Human Skill Improvement (HSI)
4. Personnel Effectiveness (PE)
5. Readiness to Change (R2C)
6. Team Building (TB)

Untuk dapat memberikan lebih dari 1 rekomendasi pelatihan, klasifikasi dilakukan per pelatihan dengan 2 kelas yaitu kelas ya dan tidak. Jadi, terjadi 6 kali proses klasifikasi yang masing-masing proses mewakili 1 pelatihan misal klasifikasi pelatihan Team Building, klasifikasi pelatihan Personnel Effectiveness dan seterusnya.

2.2 Rancangan Sistem

Untuk dapat memberikan rekomendasi pelatihan, disusun suatu sistem yang terdiri dari dua bagian seperti terlihat pada Gambar 1. Bagian yang pertama adalah bagian *Back-end* yang berfungsi untuk membangun model rekomendasi pelatihan, sedangkan bagian *Front-end* menyediakan sarana interaksi dengan pengguna untuk mendapatkan rekomendasi pelatihan dari model yang telah terbentuk.

Bagian *Back-end* menggunakan 3 buah sub sistem untuk melakukan klasifikasi. Sub sistem 1 menggunakan kombinasi algoritma PCA, diskritisasi, dan C4.5. Sub sistem 2 menggunakan kombinasi algoritma PCA dan C4.5. Sub sistem 3 menggunakan algoritma C4.5 saja. Masing-masing sub sistem akan menghasilkan model. Model tersebut kemudian dibandingkan performanya satu dengan yang lain hingga didapatkan model yang memberikan performa rekomendasi terbaik. Model dengan performa terbaik tersebut kemudian disimpan ke dalam basis data untuk digunakan dalam proses penentuan rekomendasi pelatihan secara keseluruhan yang ada pada bagian *Front-end*.



Gambar 1 Arsitektur sistem

2.2.1 Deteksi *Outlier* dengan Algoritma WAFV

Outlier adalah objek yang menyimpang secara signifikan dari objek-objek yang lain (Han et al., 2012). Teknik statistik dan *data mining* tertentu tidak dapat berfungsi dengan baik ketika terdapat *outlier* pada data yang diproses (Larose & Larose, 2015). Oleh karena itu, penghapusan *outlier* akan memberikan dampak positif dalam proses *data mining*.

Salah satu cara deteksi *outlier* adalah dengan menggunakan algoritma WAVF (*Weighted Atribut Value Frequency*). Algoritma yang dikembangkan dari algoritma AVF ini menggantikan penggunaan frekuensi nilai atribut pada algoritma AVF dengan probabilitas nilai atribut. Selain itu, metode ini juga meningkatkan performa dari metode AVF dengan mempertimbangkan tingkat kemunculan yang rendah (*sparseness*) dari tiap-

tiap atribut. Tingkat sparseness digunakan sebagai fungsi pembobotan untuk probabilitas nilai atribut. Fungsi pembobotan tersebut membuat data yang frekuensi nilai atributnya paling jarang muncul diduga kuat sebagai *outlier* (Rokhman et al., 2016).

Outlier yang telah terdeteksi kemudian dipisahkan dari data yang akan diproses. Jadi, data yang akan digunakan untuk memberikan rekomendasi atas 6 pelatihan adalah data yang sudah bersih dari *outlier*.

2.2.2 *Oversampling* dengan SMOTE

CIP (*Class Imbalance Problem*) adalah kondisi terdapat satu kelas yang ukurannya sangat kecil sekali atau sangat besar sekali jika dibandingkan dengan kelas lainnya [9]. Kondisi CIP memungkinkan *classifier* untuk mendapatkan akurasi 99% hanya dengan mengabaikan 1% kelas yang minor. Akurasi yang diperoleh akan cenderung tinggi namun kelas minor terabaikan, *classifier* cenderung mengklasifikasikan data sebagai kelas mayoritas, padahal pada beberapa kasus kelas yang menjadi perhatian adalah kelas-kelas minor. C4.5 sebagai classifier yang menggunakan konsep *entropy* akan terkendala dengan CIP karena pengukuran *entropy* hanya dapat berjalan dengan baik ketika datanya seimbang, seluruh kelas memiliki proporsi yang sama (Kishners et al., 2016).

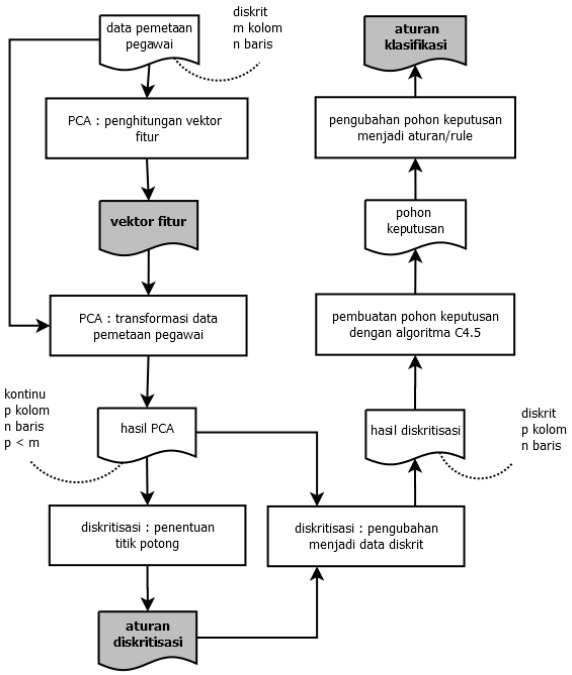
Salah satu solusi untuk mengatasi masalah CIP adalah dengan melakukan *over-sampling*. *Over-sampling* dilakukan dengan cara meningkatkan jumlah data kelas minoritas agar distribusi data menjadi seimbang (Amin et al., 2016). Teknik *over-sampling* merupakan teknik yang paling banyak digunakan dibanding teknik *under-sampling* karena teknik *under-sampling* berpotensi menghilangkan informasi penting yang ada pada kelas mayoritas (Santoso et al., 2017).

Algoritma SMOTE (*Synthetic Minority Over-sampling Technique*) mengatasi CIP dengan cara melakukan over-sampling pada kelas minoritas. *Over-sampling* yang dilakukan tidak dengan menduplikasi data kelas minor, tapi dengan membuat data sintetis berdasar data kelas minor. Pembuatan data sintetis tidak dioperasikan pada *data space* tetapi dioperasikan pada *feature space*. Pembuatan data sintetis dilakukan dengan mengambil tiap data di kelas minor kemudian membuat data sintesis disepanjang garis segmen yang melibatkan beberapa atau seluruh *k nearest neighbour* dari kelas minor (Chawla et al., 2002).

Algoritma SMOTE dapat digunakan untuk berbagai jenis tipe data. Algoritma SMOTE yang digunakan untuk tipe data kategorik adalah algoritma SMOTE Nominal atau disebut juga SMOTE-N. SMOTE-N mencari nearest neighbour dengan menggunakan *Value Difference Metric* (VDM). Pembuatan data sintetis dilakukan dengan memilih nilai fitur yang paling banyak diantara nilai fitur yang lain untuk keseluruhan suatu vektor fitur dan *k nearest neigbour*-nya (Chawla et al., 2002). Implementasi algoritma SMOTE-N pada penelitian ini dilakukan dengan menggunakan program WEKA versi 3.8.

2.2.3 Kombinasi PCA, Diskritisasi, dan C4.5

Gambar 1 menunjukkan bahwa salah satu sub sistem yang digunakan untuk menghasilkan rekomendasi pelatihan adalah sub sistem PCA, diskritisasi, dan C4.5. Sub sistem tersebut menggunakan metode yang diusulkan pada penelitian ini, PCA, diskritisasi, dan C4.5. Gambaran proses pada sub sistem tersebut dapat dilihat pada Gambar 2.



Gambar 2 Rancangan proses pada Sub Sistem 1 (PCA, diskritisasi, dan C4.5)

2.2.3.1 Algoritma PCA

Algoritma PCA pada sub sistem PCA, diskritisasi, dan C4.5 digunakan untuk mereduksi dimensi data pemetaan pegawai seperti terlihat pada Gambar 2. PCA dapat mereduksi dimensi data tanpa harus kehilangan informasi dari data asli secara signifikan sehingga data asli dengan m variabel dapat digantikan dengan k komponen sehingga dataset menjadi berukuran lebih kecil. PCA diaplikasikan pada variabel prediksi saja dengan mengabaikan variabel target (Larose dan Larose , 2015).

Dalam tahapan proses PCA akan didapatkan suatu vektor fitur. Vektor fitur ini adalah bagian dari eigen vektor yang dipilih untuk digunakan. Jika jumlah variabel data asli adalah p, dan jumlah variabel data baru adalah q, dimana dimensi data baru lebih kecil dari data asli, maka akan didapatkan vektor fitur berukuran p x q. Vektor fitur ini digunakan untuk melakukan tranformasi terhadap data asli sehingga memiliki dimensi yang lebih rendah.

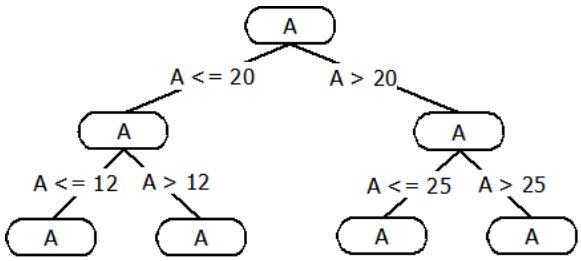
Sub sistem yang menggunakan algoritma PCA melakukan pengujian dengan menggunakan princical

component (PC) mulai dari PC ke-2 sampai dengan PC ke-13. PC yang memberikan performa terbaik akan digunakan dalam penentuan rekomendasi pelatihan.

2.2.3.2 Algoritma Diskritisasi

Diskritisasi adalah bagian dari tahap pra-pemrosesan pada *data mining* yang digunakan untuk mengubah fitur yang kontinu menjadi diskrit. Tujuan dari diskritisasi adalah untuk mengurangi jumlah kemungkinan nilai atribut dari sebuah atribut kontinu dengan cara membaginya ke dalam beberapa interval nilai berdasarkan titik potong (*cut point*) yang telah ditentukan (Kareem dan Duaimi, 2014). Penentuan titik potong dapat dilakukan sendiri oleh pengguna atau mengunakan proses perhitungan.

Diskritisasi berbasis entropi (*entropy based discretization*) adalah salah satu jenis algoritma diskritisasi terawasi yang menggunakan mekanisme *top-down*. Tujuan dari algoritma ini adalah mendapatkan partisi yang mengandung baris data dari kelas yang sama sebanyak mungkin. Entropi digunakan untuk dapat mencapai tujuan tersebut. Gambaran multi diskritisasi yang dihasilkan oleh algoritma ini dapat dilihat pada Gambar 3.



Gambar 3 Entropy Based Discretization (Fayyad dan Irani, 1993)

Data yang akan didiskritisasi diurutkan terlebih dahulu kemudian nilai yang menjadi batas dari 2 kelas dijadikan sebagai kandidat titik potong. Masing-masing kandidat titik potong dihitung *information entropy*-nya dengan menggunakan persamaan **Error! Reference source not found.**, sedangkan entropinya dihitung menggunakan persamaan **Error! Reference source not found.**. Kandidat titik potong dengan nilai *information entropy* terendah akan dipilih sebagai titik potong sehingga akan didapatkan dua buah partisi. Kedua partisi tersebut kemudian dipartisi lagi secara rekursif sampai kriteria pemberhentian tercapai (Fayyad dan Irani, 1993). Peluang algoritma ini untuk dapat meningkatkan akurasi cukup besar karena algoritma ini menggunakan informasi kelas dalam menentukan titik potong (Han et al. 2012).

$$E(A,T;S) = \frac{|S_1|}{|S|} * Ent(S_1) + \frac{|S_2|}{|S|} * Ent(S_2) \quad (1)$$

$$Ent(S) = \sum_{i=1}^{n} - p(C_i, S) * \log_2(p(C_i, S)) \quad (2)$$

S pada persamaan **Error! Reference source not found.** adalah data yang digunakan yang akan dipotong dengan titik potong T pada atribut A. S1 dan S2 adalah data dari dua interval yang menggunakan titik potong T. Ent adalah entropi dari data yang dihitung menggunakan persamaan **Error! Reference source not found.**. Pada persamaan tersebut, p(Ci,S) adalah perbandingan data sampel yang ada dalam kelas Ci dan jumlah data dalam S. n adalah jumlah kelas yang terdapat dalam S.

Kriteria pemberhentian yang digunakan pada penelitian ini adalah kriteria MDLP (*Minimum Desription Length Principle*) dan kriteria jumlah interval. Jika kriteria pemberhentiannya menggunakan jumlah interval, proses partisi akan berhenti ketika jumlah interval yang diinginkan sudah tercapai. Salah satu cara terbaik untuk menentukan jumlah interval adalah dengan menggunakan teknik Dougherty yang rumusnya dapat dilihat pada persamaan **Error! Reference source not found.** (Alvarez et al., 2013).

$$ival = \max(1, [2\log_{10}(m)]) \quad (3)$$

Pada persamaan **Error! Reference source not found.**, ival adalah jumlah estimasi interval yang merupakan nilai pembulatan ke bawah dari log m. m adalah jumlah nilai yang akan didiskritisasi. Jika nilai log m kurang dari 1, maka jumlah intervalnya adalah 1.

Selain menggunakan kriteria pemberhentian jumlah interval, kriteria pemberhentian juga dapat menggunakan kriteria MDLP yang akan menghentikan proses partisi ketika Gain(S, A) < δ (Alvarez et al., 2013). Proses partisi yang memenuhi kriteria tersebut akan ditolak. Nilai Gain(S, A) diperoleh dari persamaan **Error! Reference source not found.** sedangkan nilai δ didapatkan dari persamaan **Error! Reference source not found.** dan **Error! Reference source not found.**.

$$Gain(S, A) = Ent(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} * Ent(S_i) \quad (4)$$

$$\delta = \frac{\log_2(m-1) + \Delta(A,T;S)}{m} \quad (5)$$

$$\Delta(A,T;S) = \log_2(3^n - 2) - [n * Ent(S) - n_1 * Ent(S_1) - n_2 * Ent(S_2)] \quad (6)$$

Pada persamaan **Error! Reference source not found.**, k adalah jumlah partisi. Pada persamaan (1), m adalah jumlah data dalam S. Pada persamaan (2), n adalah jumlah kelas pada himpunan S, ni adalah jumlah kelas pada himpunan Si.

2.2.3.3 Algoritma C4.5

C4.5 adalah algoritma untuk menyelesaikan masalah klasifikasi dalam *machine learning* dan *data mining*. Algoritma yang dibuat oleh J.Ross Quinlan ini termasuk dalam jenis pohon keputusan dengan dengan model pembelajaran terawasi. Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih kriteria pemisahan (*split*) yang optimal (Larose dan Larose, 2015).

Algoritma C4.5 termasuk jenis pohon keputusan yang pembangunannya menggunakan model top-down. Langkah-langkah pembangunan pohon keputusan ini sebagai berikut (Ye, 2014):

1. Memilih *root node*
2. Menerapkan metode pemilihan pemisah (*split selection*) untuk memilih kriteria pemisah (*split criterion*) yang terbaik dan membagi data pelatihan berdasar *node*/atribut yang terpilih. Algoritma C4.5 dapat menggunakan kriteria *information gain* (atau disebut *gain*) maupun *gain ratio* tetapi *default* kriterianya adalah *gain ratio*. *Information gain*, yang rumusnya terdapat pada persamaan **Error! Reference source not found.**, adalah selisih *information entropy* sebelum dilakukan pemisahan dan sesudah dilakukan pemisahan. *Information gain* bias terhadap atribut bernilai banyak (*multivalued attribute*). Untuk mengatasi hal tersebut, C4.5 menggunakan *gain ratio* yang merupakan normalisasi dari nilai *gain* (Han et al., 2012). Rumus *gain ratio* dapat dilihat pada persamaan **Error! Reference source not found.**.

$$GainRatio(S,A) = \frac{Gain(S,A)}{\sum_{i=1}^{k} Ent(S_i)} \quad (9)$$

   Pada persamaan **Error! Reference source not found.**, k adalah jumlah partisi dalam S. Nilai Gain(S,A) pada persamaan tersebut dihitung dari persamaan (4) dan nilai Ent(Si) dihitung dengan persamaan (2).
3. Cek apakah kriteria pemberhentian sudah terpenuhi atau belum. Jika sudah terpenuhi, pembangunan pohon keputusan akan dihentikan. Jika tidak terpenuhi, ulangi kembali langkah ke-2 dengan memilih sebuah node untuk pemisahan.

Pemberhentian dengan menggunakan kriteria pemberhentian berdasar homogenitas data akan dilakukan ketika tiap *leaf node* memiliki data yang homogen. Data disebut homogen ketika seluruh data pada *leaf node* tersebut mempunyai nilai target yang sama.

2.3 Hasil dan Pembahasan

Untuk mendapatkan rekomendasi pelatihan yang terbaik, dilakukan pembandingan performa kombinasi algoritma PCA, diskritisasi, dan C4.5 terhadap algoritma C4.5 saja dan kombinasi algoritma C4.5 dan PCA. Setelah dilakukan pengujian dengan menggunakan skema pengujian 10-*fold cross validation*, metode terbaik didapatkan dan hasil pengukuran performa untuk masing-masing pelatihan dapat dilihat pada Tabel 1.

Tabel 1 Pelatihan dan metode rekomendasinya

NO	Pelatihan	Meth	Aku	Pre	Rec
1	AMT	C.b	0.692	0.676	0.597
2	ECS	C.a	0.845	0.813	0.859
3	HIS	C.a	0.860	0.852	0.852
4	PE	C.a	0.645	0.650	0.745
5	R2C	C.b	0.904	0.895	0.915
6	TB	C.a	0.566	0.596	0.416

Keterangan: Meth: Metode, Aku: akurasi, Pre: presisi, Rec: Recall, C.a: kombinasi algoritma PCA, diskritisasi dengan kriteria pemberhentian jumlah interval, dan C4.5, C.b: kombinasi algoritma PCA, diskritisasi dengan kriteria pemberhentian MDLP, dan C4.5

Dari Tabel 1 di atas terlihat bahwa 5 dari 6 rekomendasi pelatihan diperoleh dari metode kombinasi PCA, diskritisasi, dan C4.5. Pelatihan AMT dan pelatihan *Readiness to Change* rekomendasi terbaiknya diperoleh dari metode tersebut dengan kriteria pemberhentian MDLP. Pelatihan *Effective Communication Skill*, *Human Skill Improvement*, dan *Personnel Effectiveness* rekomendasi terbaiknya diperoleh dari metode PCA, diskritisasi, dan C4.5 dengan kriteria pemberhentian jumlah interval. Pelatihan *Team Building* rekomendasi terbaiknya diperoleh dari metode PCA, dan C4.5.

Tabel 2 menunjukkan contoh hasil rekomendasi pelatihan yang diberikan oleh sistem. Sebagai pembanding, rekomendasi pelatihan yang diberikan assessor juga ditampilkan. Akurasi rekomendasi yang diberikan oleh sistem dihitung per pegawai. Akurasi rekomendasi untuk data yang ada pada dataset dilakukan dengan cara menghitung rata-rata dari akurasi rekomendasi per pegawai yang ada pada dataset tersebut.

Tabel 2 Hasil rekomendasi pelatihan

NO	Pegawai	Rekomendasi Pelatihan		Akurasi (%)
		Assesor	Model	
1	Pegawai no. 047	EFS PE TB AMT	EFS PE TB	83,3
2	Pegawai no. 141	PE	PE	100
3	Pegawai no. 011	PE AMT	PE	83,3
4	Pegawai no. 068	PE TB	PE	83,3
5	Pegawai no. 049	HSI PE TB	HSI PE	83,3
6	Pegawai no. 114	EFS PE TB	EFS PE	83,3
7	Pegawai no. 043	EFS	EFS	100
8	Pegawai no. 140	EFS PE TB	EFS PE AMT	66,7

9	Pegawai no. 028	AMT TB	AMT PE	66,7
10	Pegawai no. 041	PE TB	PE	83,3

Banyak terjadi kesalahan rekomendasi pelatihan Team Building seperti terlihat pada Tabel 2. Hal tersebut dikarenakan rendahnya performa rekomendasi Team Building yang akurasinya hanya sekitar 53,%. Setelah diteliti lebih lanjut, rendahnya performa rekomendasi untuk pelatihan *Team Building* terjadi karena adanya ketidakkonsistenan data. Sebagai contoh, ada pegawai yang nilai aspek psikologisnya dominan pada level 3 dan direkomendasikan untuk mengikuti pelatihan *Team Building*. Sementara itu, ada pegawai lain yang nilai aspek psikologisnya lebih rendah dan dominan di level 1 malah tidak direkomendasikan mengikuti pelatihan *Team Building*. Pegawai tersebut nilai aspek kerjasamanya juga berada di level 1. Nilai aspek kerjasama menjadi salah satu nilai terpenting dalam menentukan perlu tidaknya pelatihan *Team Building*. Pegawai dengan nilai kerjasama yang rendah memerlukan pelatihan *Team Building*. Jika pegawai dengan nilai aspek kerjasama adalah 3 dianggap perlu mengikuti pelatihan *Team Building*, seharusnya pegawai dengan nilai aspek kerjasama di bawah 3 lebih perlu mengikuti pelatihan *Team Building*.

3. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa penentuan rekomendasi pelatihan pengembangan diri bagi pegawai dapat dilakukan menggunakan algoritma C4.5 yang dikombinasikan dengan PCA dan diskritisasi dengan tingkat akurasi rata-rata adalah 86,6%. Pendekatan yang diusulkan dalam penelitian ini yaitu dengan mengkombinasikan algoritma PCA, diskritisasi, dan algoritma C4.5. Hasil dari kombinasi metode tersebut menunjukkan performa yang lebih baik daripada menggunakan kombinasi algoritma C4.5 dan PCA, serta algoritma C4.5 saja untuk kasus penentuan rekomendasi pelatihan pengembangan diri. Hal tersebut terbukti dengan didapatkannya rekomendasi terbaik untuk seluruh jenis pelatihan dari metode PCA, diskritisasi, dan C4.5.

Daftar Pustaka

Alvarez, M.A., Carrasco, J.A. & Martinez, J.F., 2013. Combining Techniques to Find the Number of Bins for Discretization. In *32nd International Conference of the Chilean Computer Science Society*. Temuco, pp. 54–57.

Hermanto, B., SN, A., 2017, Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5, IJCCS (Indonesian J. Comput. Cybern. Syst), vol. 11, no. 1, p. 43 [Online]. Available: https://journal.ugm.ac.id/ijccs/article/ view/15946. [Accessed: 11-Sep-2017]

Amin, A. et al., 2016. Comparing Oversampling Techniques to Handle the Class Imbalance Problem : A Customer Churn Prediction Case Study. , 4(Ml).

Chawla, N. V et al., 2002. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp.321–357.

Fayyad, U.M. & Irani, K.B., 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Proceddings of 13th International Conference on Artificial Inatelligence*. pp. 1022–1027. Available at: http://www.decom.ufop.br/luiz/site_media/uploads/arquivos/bcc444_pcc142/multiintervaldiscretizationofcontinuousvaluedattributesforclassificationlearning1993.pdf.

Han, J., Kamber, M. & Pei, J., 2012. *Data Mining: Concepts and Techniques* 3rd ed., San Francisco: Morgan Kaufmann Publishers.

Hussain, A., Rao, M.K. & Mahmood, A.M., 2013. An Optimized Approach To Generate Simplified Decision Trees. In *IEEE International Conference on Computational Intelligence and Computing Research*. Tamilnadu: IEEE.

Jantan, H., Hamdan, A.R. & Othman, Z.A., 2011. Talent Knowledge Acquisition using Data Mining Classification Techniques. In *Conference on Data Mining and Optimization*. Selangor, pp. 32–37.

Kareem, I.A. & Duaimi, M.G., 2014. Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization. *International Journal of Computer Science and Mobile Computing*, 3(6), pp.176–183.

Larose, D.T. & Larose, C.D., 2015. *Data Mining And Predictive Analytics* 2nd ed., New Jersey: John Wiley & Sons, Inc.

Rokhman, N., Winarko, E. & Subanar, 2016. Improving the Performance of Outlier Detection Methods for Categorical Data By Using Weighting Function. *Journal of Theoretical and Applied Information Technology*, 83(3), pp.327–336.

Kishners A., Parshutin S., Gorskis H., 2016, Entropy-Based Classifier Enhancement to Handle Imbalanced Class Problem, ICTE 2016, Latvia, p. 588, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917301771. [Accessed: 11-Sep-2017]

Santoso, B., Wijayanto, H., Notodipuro, K.A., Sartono, B., 2017. Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review. In *58 th IOP Conference Series: Earth and Environmental Science*. IOP Publishing, pp. 1–8.

Strohmeier, S. & Piazza, F., 2013. Domain Driven Data Mining in Human Resource Management: A Review of Current Research. *Expert Systems with Applications*, 40(7), pp.2410–2420. Available at: http://dx.doi.org/10.1016/j.eswa.2012.10.059.

Ye, N., 2014. *Data mining*, Boca Raton: CRC Press.